

Survey of Numerical Techniques for Grouping

TOM BERGAN

Kaptein W. Wilhelmsen og Frues Bakteriologiske Institutt, Rikshospitalet, University of Oslo, Oslo, Norway

INTRODUCTION.....	379
METRICS OF SIMILARITY.....	379
Coefficients of Association.....	379
Distance Coefficients.....	380
Angular Coefficients.....	381
Correlation Coefficients.....	381
Weighting of Characters.....	381
Choice of Similarity Index Attributes.....	382
CLASSIFICATION PROCEDURES.....	382
General Remarks.....	382
Agglomerative Techniques.....	382
Divisive Techniques.....	386
Miscellaneous Techniques.....	386
Controls on the Results of Classification.....	387
CONCLUDING REMARKS.....	387
LITERATURE CITED.....	387

INTRODUCTION

Numerical grouping procedures have been employed in various scientific fields. Best known to bacteriologists would be numerical taxonomy (14, 28, 38, 75, 78), where a particularly wide range of different approaches has been investigated. Otherwise, these procedures have been employed in psychology to demonstrate character associations (71, 72), in ecology (39, 87), and in anthropology (51). Recently, numerical grouping procedures have been introduced, e.g., in linguistics (58), in archaeology (36), and in medicine as an aid to the classification and diagnosis of disease (3, 6, 33). Such techniques have demonstrated association between points in the second dimension such as (i) in the dispersion pattern of chromosomes (17), or (ii) within technology, e.g., for interconnecting towns with the shortest possible mileage of telephone cable (62).

Numerical grouping proceeds in two steps. First, metrics of similarity are extracted to express the degree of association between the items considered. Subsequently, a sorting strategy is applied to the similarity metrics to achieve the grouping as such. Techniques now available for numerical grouping have increased in number and become more sophisticated during the last decade. To disclose the rationale for selecting specific procedures for analysis of bacteriophage data (Bergan, *manuscripts in preparation*), a brief presentation of the many alternative procedures currently available is provided. The emphasis in this paper is on numerical grouping procedures

developed since the extensive survey of Sokal and Sneath (81). However, for the sake of completeness, the earlier methods are also discussed. Other recent surveys of aspects of the related literature are found in Ball (1), in Cole (11), and in Harmann (32).

Grouping of bacteriophages by numerical procedures according to their lytic spectra will be presented in later reports (Bergan, *manuscripts in preparation*).

METRICS OF SIMILARITY

Coefficients of Association

The various approaches for computing the resemblance between "operational numerical units" may conveniently be subdivided into four categories: (i) coefficients of association, (ii) coefficients of distance, (iii) angular coefficients, and (iv) coefficients of correlation. Collectively, they are denoted coefficients, or indices, of resemblance, similarity, agreement, or congruity. [In this paper, an item considered for grouping will be referred to as an operational numerical unit (ONU) analogous to the more connotatively restricted term "operational taxonomic unit" (OTU) of Sokal and Sneath (81).]

The most commonly encountered metrics in taxonomy have been (i) the "Jaccard-Sneath index" (40, 73, 81)

$$S_{JB} = \frac{n_{JK}}{n_{JK} + n_{Jk} + n_{jK}}$$

and (ii) the "simple matching coefficient" (79):

$$S_{SM} = \frac{n_{JK} + n_{jK}}{n}$$

where, in accordance with Sokal and Sneath (81), n is the total number of characteristics used for comparison, n_{JK} denotes all positive characters in common for both ONU's, n_{jK} means the number of characters where ONU "J" has positive and ONU "K" negative reactions, and vice versa for n_{jK} , and n_{jK} signifies the number of negative characters shared.

Other coefficients of association can be constructed by appropriate variations in the enumeration of negative and positive matches, in their weighting, or by including the sums of rows and columns for each ONU pair 2×2 table. Several alternatives have been described by Cole (12), Sokal, and Sneath (81), Ball (1), and Hall (31).

The differences in S_{JS} and S_{SM} have evolved from differing philosophies on the significance of negative matches. In taxonomy, events of conjoint character absence (nonoccurrence) are often considered just as important as positive matches (81). Evidence has accumulated to show that inclusion of negative matches may produce sharper group demarcation (4, 34). Sneath (73) stated his own rationale for excluding negative similarities to be that he considered the class of negative properties almost indefinite; however, this has been challenged (5). Coefficients counting both minus and plus are more resistant to change in value upon further addition of characters. Certainly, employing a large number of attributes leads to stability in similarity coefficient values. Sokal and Sneath (81) suggested that 60 or more characters would be necessary.

A comparison of various other similarity metrics appears in reference 81, but several indices have been introduced more recently. McQuitty (55) used an "index of association" which was the simple sum of characteristics shared by two individuals. Williams et al. (88) mentioned a "nonmetric coefficient":

$$S_{WLL} = \frac{n_{JK} + n_{jK}}{2n_{JK} + n_{jK} + n_{jK}}$$

Since only nonmatches appeared in the numerator, this was a divisive metric, i.e., measured dissimilarity rather than affinity. The double counting of positive matches in the denominator was based on the rationale that it actually contained two positive reactions such that the weighting of reactions became the same regardless of equality or difference for the reaction (character). The

denominator equals that in Sørensen's index (82) [the "coincidence index" of Dice (15)] and that of an unnamed index presented by Sokal and Sneath (81).

Johnson (42) used the formula:

$$S(xy) = \frac{n_{JK}}{n_{JK} + n_{jK}} + \frac{n_{jK}}{n_{JK} + n_{jK}}$$

which has twice the numerical value of the Kulczynski coefficient (44, 81).

Hubálek (38) has constructed a similarity index intended for multivariate characters

$$S_H = \frac{r \sum p - \sum |d|}{r \sum p}$$

where r is the scoring scale range, $\sum p$ is the number of concordant characters for the ONU pair considered, and $\sum |d|$ is the absolute total difference for all characters. This index applies only when all variables possess identical score ranges.

In addition to the similarity index, Lance and Williams (45, 88) also computed an "information statistic": $I = n \cdot H$, where n is total number, H is a "system entropy":

$$H = - \sum_{s=1}^u p_s \cdot \log p_s$$

in which p_s is the probability of the s -th state over which the system may vary. In principle, the I could be computed for any pair of items, but it is not defined for continuous numerical data. Furthermore, for u bivariate characters it reduces to $I = 2(n_{jK} + n_{jK}) \log 2$ (88).

Distance Coefficients

The distance coefficients have been exhaustively described by Sokal and Sneath (81). The most commonly used distance statistic is the simple mean distance (\bar{d}_{ij}) between ONU's "I" and "J" for all characters, κ :

$$\bar{d}_{ij}^2 = \frac{1}{n} \sum_{\kappa=1}^u (x_{\kappa i} - x_{\kappa j})^2$$

Coefficients of association have been transformed by various measures to express dissimilarity in expressions subsequently treated as distances:

$$d = (1 - S_{SM})^{\frac{1}{2}} \quad (80)$$

$$d = [2(1 - S_{ij})]^{\frac{1}{2}} \quad (27)$$

$$d = -\log_2 S_{RT} \quad (65)$$

$$d = \log_2 \frac{n_{JK} + n_{JK}}{n} + n_{JK} \quad (5)$$

When the coefficient of the simple average Euclidean distance is applied to bivariate characters, the result is the expression:

$$d^2 = \frac{n_{JK} + n_{JK}}{n}$$

which is equivalent to $d^2 = (1 - S_{SM})$.

Angular Coefficients

Angular coefficients are conceptually related to distance coefficients. Both consider a geometrical model with the classifiable items distributed in a multidimensional character space. The angular coefficients describe the angle between lines drawn from each single object to the character space origin. The angle may be measured by its cosine and then transformed. Bhattacharyya (6) used the angle ϕ itself; $\cos \phi$ may be transformed to a distance:

$$d = (2 - 2\cos\phi)^{\frac{1}{2}}.$$

Boyce (Ph.D. thesis, Univ. of Oxford, 1965) used the cosine value without transformation:

$$\cos\phi_{ij} = \frac{\sum_{k=1}^n x_{ki} x_{kj}}{\left(\sum_{k=1}^n x_{ki}^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^n x_{kj}^2\right)^{\frac{1}{2}}}$$

where x_{ki} is the value of character k for item "I," x_{kj} for item "J," and n is the total number of characteristics.

Correlation Coefficients

Among the correlation coefficients, the common product moment correlation coefficient (r) defined in any standard textbook of statistics (86) appears to be the most suitable (56, 67, 79, 81). The Kendall rank correlation coefficient has also been employed (22, 23). It can be shown (Bergan, *manuscript in preparation*) that the product moment coefficient for a 2×2 table transforms to the Yule ϕ -coefficient, which is listed by Sokal and Sneath (81) as a coefficient of association and has been used elsewhere for numerical allocation (8; Bergan, *manuscript in preparation*).

χ^2 For a 2×2 contingency table may also be used as a similarity statistic (16, 23, 81). It has been pointed out that, for n items, the $\chi^2 = n \cdot r^2$ (81).

Weighting of Characters

The principle of weighting is of key significance in numerical taxonomy, where the Adansonian

principle of assigning equal weight to all characters has been an axiom. This is opposed to the Aristotelian idea that in classification certain characters have greater a priori information content than others (54). Unequal significance of characters in traditional taxonomy has been an organic part of the subjective skill of hierarchy construction. Unavoidably containing an arbitrary element, traditional procedures are somewhat opposed to the very principle of objectivity in numerical techniques.

The weighting procedure, however, can be made objective such that concordance of rare traits contributes more to similarity than identity in common variates. With the assumption that attributes are stochastic elements, Goodall (23, 24) used a "probabilistic similarity index," where the cumulated probability of an observed pair was calculated either exactly or by the χ^2 approximation based on an ONU pair contingency table. The similarity index, S_{Goodall} , for a pair of individuals was the complement of the combined cumulative probability of their attributes (p_i):

$$S_{\text{Goodall}} = 1 - \sum_{i=1}^u p_i$$

Gower (*unpublished data*) intuitively has followed a similar approach of probabilistic character weighting:

$$S_{\text{Gower}} = \frac{\sum_{k=1}^u s_{kij}}{\sum_{k=1}^u w_{kij}}$$

where s_{kij} is a score and w_{kij} is a weight for each character k . The $w_{kij} = 1$ for valid comparisons; otherwise $w_{kij} = 0$. The $s_{kij} = 0$ when $w_{kij} = 0$. In the Gower coefficient, the numerator is calculated differently for bivariate and multivariate characters. For dichotomies, $S_{kij} = n_{JK}$ and consequently, $S_{\text{Gower}} = S_{JK}$. Metric variables are standardized by the formula:

$$S = \frac{1 - |x_i - x_j|}{R_k}$$

where x -subscripts are the attribute values for each ONU pair and R_k is the total range of values for the character k .

Similar principles were followed by Baron and Fraser (3) in disease classification. They considered infrequent symptoms and signs to carry a higher information value. Avoiding any a priori or biased weighting, this becomes an acceptable approach. Any other presently conceivable principle for character weighting will inadvertently cause misleading results. In taxonomy, weighting in advance is unsatisfactory,

also because it would be self-contradictory or impossible for organisms which hitherto have been either unknown or are unrelated to any known taxon.

Related to the problem of character weighting is the choice of a suitable coding procedure for the observational data. Characters existing in only two mutually exclusive states present no problem in this regard, but care must be exercised in transforming traits which exist in several different states or which are a composite of multiple subtraits. In such situations, an element of character weighting may inadvertently be introduced, unless special precautions are taken and the problem is defined properly. In a recent presentation by Lockhart (48, 50), the reader will find an excellent discussion on the proper coding of data for computer use.

Choice of Similarity Index Attributes

The choice of the proper similarity statistic in the end is left largely to individual preference. Because of their different properties, the indices decisively influence the similarity pattern and the ensuing clustering results. Accordingly, similarity metrics by and large are nonmonotonically related; it is not predictable that for three given ONU's, "*H*," "*I*," and "*J*," the similarity S_{hj} between "*H*" and "*J*" is always larger than for "*I*" and "*J*," S_{ij} , for any index.

Qualitative data are handled well by coefficients of association, whereas metric data are eminently treated by coefficients of correlation or distance, S_{Goodall} and S_{Gower} . For metric data, distance has much to commend it (27), particularly when variables are expressed in a standardized, nondimensional scale to avoid aberrant influence from size (78). This may, for instance, be achieved by dividing each measurement by the standard deviation for all observations of that attribute, or by coding between the limits 0 and 1 (9). Standardization is superfluous (i) when all variates are measured in the same units, and (ii) for presence-absence data (81).

In taxonomy, the proper selection of attributes is decisive. This is mentioned here for its implication in other types of problems, notably the bacteriophage study to be presented (Bergan, *manuscripts in preparation*). The characters should represent an exhaustive spectrum within the character sphere.

For part of the bacteriophage grouping to be reported, S_{JS} and S_{SM} were selected, since each represents opposites in the philosophy of the importance of negative and positive matches and thus may be considered to represent a synthesis of many of the index existing for bivariate characters.

CLASSIFICATION PROCEDURES

General Remarks

Numerical allocating procedures are divisible in two broad categories: (i) clustering procedures, and (ii) techniques based on vector calculus (e.g., factor analysis and principal components analysis). Clustering may be agglomerative or divisive. The former starts with a single ONU and brings about progressive ONU fusion; the latter initially considers the entire population as one unit and progressively subdivides it according to single features. Divisive procedures are frequently described as monothetic, but in reality they are usually oligothetic (87) and are, in my opinion, more useful for constructing diagnostic keys than for demonstrating hierarchical structure.

A basic requirement for objective procedures is that they be defined in detail. The same similarity matrix provided different scientists must obtain identical classification patterns. A few of the procedures dealt with below actually contain provisions contrary to this fundamental prerequisite.

It is pertinent to note that, owing to the tendency for numerical grouping procedures to involve considerable generalization, conclusions should be made with some reservation. According to individual preference and the nature of the grouping problem, a series of procedures is available for objective allocation.

Agglomerative Techniques

The crudest means of creating structure in a Q matrix (relationship of elements; R matrix analysis for relationships of characters) is by "differential shading" (46, 81). This involves the repetitious rearrangement of rows and columns to create unified areas of high similarity. This method, however, is unwieldy for large matrices where each square, in addition, becomes small and optic differentiation consequently becomes inadequate.

Some workers prefer a "bar diagram" to show ONU affinities (46). This appears as a histogram with each ONU along the abscissa and the per cent similarity along the ordinate. One starts nearest the ordinate with the two elements which have the highest affinity and draws a line horizontally at the per cent similarity level involved. The third element to be entered is that which has the highest affinity to either of the first two ONU's; again, a horizontal line indicates the similarity level. The remaining ONU's are entered one by one according to their highest affinity with the foregoing ONU. By connecting the horizontal lines with vertical ones, one obtains a simple representation of intragroup similarity. The procedure, though, hides the nuances of intragroup affinity and seems most suitable for rela-

tively homogeneous groups or a small number of ONU's.

More sophisticated are the "clustering procedures" which constitute a large class of variant approaches to the grouping problem. These usually require the aid of electronic data processing (EDP) as has been outlined briefly by Quadling (64). Rubin (68) presented theoretical aspects of clustering and suggested a "hill climbing algorithm" as an approach to the programming of such procedures. The short survey by Quadling (64) gives the necessary background information needed by the biologist. In most instances, he would prefer to seek the highly specialized services of a trained EDP programmer rather than attempting to solve this unusual task by himself. However, the biologist needs a knowledge of mathematical details and objectives of the clustering procedures to make a suitable choice of procedures.

Various cluster procedures have been eminently described by Sokal and Sneath (81). The "elementary cluster analysis" consists of arbitrarily selecting successively diminishing similarity levels above which subsequent OTU pair similarities are scrutinized. This fails to render a satisfactory structure except for small matrices.

Objective procedures are inherent in Sneath's (73) "clustering by single linkage." Here an ONU is admitted into the cluster containing the (possibly one) other individual with which the ONU has the highest linkage. Since no condition exists for overall affinity within clusters, these become progressively heterogeneous. Single linkage, therefore, has also been labeled a "chaining procedure." Clustering by single linkage is related to the "ramifying linkage method" and the "approximate delimitation method" (10).

McQuitty has approached the linkage analysis problem in various ways and developed a "hierarchical linkage analysis" (55) which seems essentially similar to single linkage cluster analysis.

The Sørensen's "complete linkage" (82) admits that ONU which has the highest similarity to every member already in the cluster. This procedure is also called the "nearest neighbour linkage" method (45). Complete linkage and "highest linkage" entail irrelevant grouping bias for overlapping clusters, and they are therefore surpassed by the "clustering by average linkage" elaborated by Sokal and Michener (79).

The "complete linkage" is not to be confused with the procedure of Nigel da Silva and Holt (61) where the "highest linkage" criterion for group formation is used (20). This procedure joins two clusters at the highest similarity level found between any ONU of the first group with any ONU of the second group.

Average linkage procedures are involved in an entire class of clustering techniques. They base union of any individual ONU ("pair group" method) or several ONU's ("variable group" method) to any other particular ONU or cluster on the basis of the average similarity of the potential entrant with the previous members of the cluster considered. Those particular ONU's or clusters may fuse, which results in the lowest possible drop in the recalculated average correlation index. As clusters enlarge they become increasingly heterogeneous, and progressively more remote relatives are admitted; the value of the average cluster similarity simultaneously becomes reduced. In the variable group method, the permissible similarity drop for each cluster cycle is defined in advance. The level 0.03 has been found acceptable for S_{SM} (79) and 0.20 for a correlation coefficient (57), but this value would vary with the kind of grouping problem, the matrix size, the particular coefficient of congruity used, and the degree of overall ONU set homogeneity. Accordingly, the variable group procedure involves a subjective element. The pair group method takes longer to calculate but gives more detail in branching, and consequently it seems preferable for most problems. Boyce (Ph.D. thesis, Univ. of Oxford, 1965) regarded the pair group method as easier to programme for computers than the variable group method.

Recalculation of affinities after each clustering cycle in the pair group average clustering procedures could be achieved by the Spearman sums of variables formula used initially (79). This has now been abandoned, owing to the occasional occurrence of reversals in the level of correlation (81), in that a similarity S_{ij} for ONU's "I" and "J" could be lower than the cluster similarity $S_{(ij)k}$ after addition to the cluster "I + U" of another ONU, "K" ($S_{ij} < S_{(ij)k}$). In any case, Spearman's method is suitable only for correlation coefficients; application to distances or coefficients of association is not proper. For these reasons, new affinities are now calculated as arithmetic averages of all coefficients involved in the prospective union of any two clusters (74, 80; F. J. Rohlf, Ph.D. thesis, Univ. of Kansas, 1963). Recalculation of affinities during each clustering cycle may proceed by an "unweighted linkage" (UWPGA) or a "weighted linkage" (WPGA) procedure. Both methods have in common the circumstance that clustering starts with the ONU's with the highest calculated similarities. In the unweighted procedure (UWPGA), each individual ONU has equal weight. For two clusters of 3 ("A," "B," "C") and 2 ("D," "E") elements, the average similarity equals the sum of similarities between the ONU's of the two clusters divided by the number of in-

dexes. For UWPGA, regardless of the internal cluster structure, the average similarity for the two clusters will be:

$$S_{(abc)de} = \frac{S_{ad} + S_{ae} + S_{bd} + S_{be} + S_{cd} + S_{ce}}{6}$$

where, e.g., S_{ad} denotes the similarity between "A" and "D".

For the unweighted clustering procedure, reference is always made to the original similarity matrix. With the weighted clustering procedure (WPGA), on the contrary, a new similarity matrix is recalculated after each cycle. An example of the calculations involved in the WPGA is demonstrated in Table 1.

With WPGA, the ONU's clustered during the first clustering cycle contribute less to the average similarity figure than ONU's admitted more recently.

Sokal and Sneath (81, p. 190-191) quoted their reason for adopting the weighted approach to be that "underlying assumed phylogenetic causes of the phenetic relationships under study" made it less objectionable. Rationale for favoring the weighted method is also given by Gower (27): "These [individuals] would be represented by identical points and we would not want them to bias by sheer weight of numbers the cluster they would inevitably form." Thus, WPGA will somewhat counteract any over-representation of one (or a few) type(s) of ONU('s). Weighting was likewise preferred by Kendrick and Proctor (43). However, in recent publications, UWPGA hierarchies have repeatedly been more closely cor-

related with the similarity matrices (76; Bergan, *manuscripts in preparation*; Boyce, Ph.D. thesis, Univ. of Oxford, 1963).

Gower (27) suggested a "centroid average" clustering approach. Various other schemes similar in principle to the Sokal-Michener method, but certainly different in computational detail, have been formulated. Ball (1) subdivided the important procedures into: (i) probabilistic, (ii) clustering, and (iii) clumping techniques. With Hall he developed ISODATA procedure where grouping was based on decision theory. This estimated the probability of pattern occurrence. Owing to the large computer storage and computation time involved, the approach was considered impractical for all but small matrices. The distinction made (1) between clustering and clumping techniques seems to render recognition of mere details.

Boyce (Ph.D. thesis, Univ. of Oxford, 1963) has described an alternative procedure, "centroid cluster analysis" (CCA) cognate to the group average procedure. CCA is applicable only to distances; for affinity or angular coefficients, consequently, transformation to distance (actually dissimilarities) is necessary. Boyce used the formula:

$$d_{ij}^2 = 2(1 - a_{ij})$$

where a_{ij} was the cosine or the angular correlation coefficient between two ONU's. Accordingly, distances between centroids were expressed by

$$d_c^2 = \bar{B} - \frac{t_1 - 1}{2t_1} \bar{W}_1 - \frac{t_2 - 1}{2t_2} \bar{W}_2$$

TABLE 1. Calculation of the weighted-pair group-clustering procedure with simple averages (WPGA) using an arbitrarily selected similarity matrix

Matrix	ONU	ONU					Equations
		A	B	C	D	E	
First	A	x					$A' = A + B$ $D' = D + E$
	B	95	x				
	C	35	5	x			
	D	80	70	10	x		
	E	20	30	40	90	x	
Second ^a	A'	A'	C	D'			$A'' = A' + D'$
	C	x					
	D'	20	x				
		50	25	x			
Third	A''	A''	C				
		x					
	C	22.5	x				

^a The formulas used for two of the calculations are: $S_{A'D'} = (S_{Ad} + S_{Ae} + S_{bd} + S_{be})/4$ and $S_{A''C} = (S_{A'e} + S_{d'e})/2$. S_{ad} signifies the similarity index between element "A" and element "D".

where \bar{B} is the mean of the squared distances between the t_1 members of the first group and the t_2 members of the second, \bar{W}_1 is the mean of the $\frac{1}{2}t_1(t_1 - 1)$ squared distances within the first and \bar{W}_2 within the second group. The unweighted centroid cluster analysis was less accurate than the UWPGA.

Lance and Williams (45) introduced the "group average procedure" wherein the average was based upon the sum of similarities for all pairs of individuals from each of the groups considered. Wishart (89) has shown that this is equivalent to the minimization of $s_i^2 + s_j^2 + d_{ij}^2$ where s_i^2 and s_j^2 are the variances of items "I" and "J" and d_{ij}^2 is the squared distance between each of their centroids. If an association coefficient is employed, d_{ij}^2 is substituted by S_{ij} .

Working with his probability similarity index, Goodall (23) used a Lancaster χ^2 approximation during clustering. However, the calculation involved with large matrices quickly becomes insurmountable. Goodall (25), in addition, introduced hypothesis testing in cluster analysis. This started with the null hypothesis that the set of individuals considered formed, or sampled, a single population. Only if the null hypothesis had to be rejected at the significance level chosen, should one proceed to subdivide the set into classes. Lance and Williams (45), after extensive testing of various clustering techniques, formulated a "generalized sorting strategy" by the linear expression:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

By changes in parameter values, this equation described all major agglomerative sorting procedures (furthest linkage, complete linkage, median linkage, centroid linkage, and average linkage in Lance and Williams' nomenclature). In "furthest linkage," the distance between the most distant pair of ONU's in each cluster was measured. Reversal of fusion level was a problem for "centroid linkage" (88). Average group linkage was obtained when:

$$\alpha_i = n_i/n_k; \alpha_j = n_j/n_k; \beta = \gamma = 0$$

By setting

$$\alpha_i + \alpha_j + \beta = 1; \alpha_i = \alpha_j; \beta < 1; \gamma = 0$$

and giving β steadily decreasing values below 1 (including negative values), sets of hierarchies with increasingly intense clustering may be obtained. *t'* Mannetje (52, 53) found that for *Rhizobium* the most meaningful results were obtained when $\alpha_i = \alpha_j = 0.625; \beta = -0.25; \gamma = 0$.

Centroid procedures with a random selection of "starter" individuals have been employed by Ball

and Hall (2), McQueen (50), and Sebestyen (70). The starting points fused into one single centroid, and further ONU's were added by various procedures based on distance.

Johnson (42) described two "hierarchical clustering schemes" (HCS), the "minimum method" and the "maximum method." These are essentially similar to the single linkage and the complete linkage clustering procedures. Since the similarity matrix for each clustering cycle is recomputed directly from the preceding matrix and not from the original matrix, the HCS becomes a weighted procedure. To maintain the sequence of monotone transformations of similarity values, the HCS was based on each individual S value, not on averages as has previously been found superior (79, 81). It was asserted that the HCS methods were conceptually simpler and that the clear-cut meaning of the "connected" or "compact" solutions (42) would otherwise be lost, arguments which are contradictory to other presentations (81).

The "central clustering procedure" (or "nodal clustering") of Rogers and Tanimoto (65) differ markedly from the preceding techniques. Their cluster analysis was designed for taxonomists who wanted an aid in classification but desired to retain the elements of skill and art intrinsic to traditional taxonomy. The central clustering (65) has a built-in feedback from machine to man at all steps in the procedure and amply allows for the operator's subjective judgement. Thus, although mathematically refined, and regardless of its achievements in certain instances, this procedure is not objective to the extent desirable for numerical allocation. Computational details appear elsewhere (65, 81). In principle, the central clustering commences by calculating an R_i index which for each ONU signifies the number of paired comparisons where at least one character is shared, i.e.,

$$S_{RT} = \frac{n_{jk} + n_{ik}}{n + n_{jk} + n_{ik}} > 0$$

and selecting the ONU with the highest R_i as centroid. For this, the sums of all $n - 1$ distances are calculated: $H_{io} = \sum d_{ij}$, where $d_{ij} = -\log_2 S_{RT}$. The successive admittance of nodes is regulated by a "measure of inhomogeneity," $u_n(d_{ij})$. The central clustering procedure is inadequate because it merely delineates primary nodes and does not indicate how they are inter-related. Colman (13) found the results obtained by the above procedure confusing and inferior even to single linkage. Related to the central clustering is the "cluster similarity cluster analysis" (19) which coordinates concepts employed by several other procedures. The distance index

calculated by Rogers and Tanimoto (65) was replaced by a "cluster similarity figure."

Bonner (7) described a "method III" which, as explained on the hand of distances (89), appears to be fairly close to Sokal and Michener's methods (79, 81), except for its commencing with a chance selection of an ONU. In the Hyvärinen (39) approach, the ONU's were clustered according to an "information loss" entropy after the "most typical items" had been identified. Inherent in the clusters formed was a grave diameter constraint similar to that found in the complete linkage procedure (89).

Jancey (41) evaluated an interesting approach whereby k points in space were randomly selected. The ONU's were then allocated to their nearest random center (class point) which, from the start, was not necessarily inhabited by an ONU. The center of group gravity around each point was then calculated as the mean of the coordinates of the group members on each axis, and, subsequently, the class point moved to the center of gravity. This process was repeated until no further shift occurred for the point of class gravity. Optimal grouping was reached through the meticulous process of trying different starters and accepting the situation with the least within-group variance. Akin to this procedure is another technique described by Forgey (21).

Wishart has described a "mode analysis" aimed at stabilizing procedures like the two preceding ones. Although appearing suitable for special problems, i.e., grouping star clusters (88), its utility in taxonomy and related fields remains unsettled.

Divisive Techniques

Divisive techniques are for instance (i) the "monothetic clustering technique" (49) where a "cumulative difference" (d_c) indicates branching, and (ii) the procedure of Edwards and Cavalli-Sforza (17). The last procedure employed squared distances and achieved optimal division when the intergroup sum of squares was the largest and simultaneously the intragroup sum of squares was the smallest possible. Cluster density was characterized by its variance, i.e., the ratio of the intra-cluster sum of squares to the number of ONU's in it. To this end, all possible intergroup and intragroup sums of squares were calculated for each unit, a detail which seriously limited the usefulness of the method. Computers with 5-sec access time need $(n-1)2^{n-1}$ sec computing time to examine the $(2^{n-1} - 1)$ possible splits. Thus, the task quickly becomes insurmountable. One hundred hours are needed for as few as 21 ONU's. For the 113-element Q-matrix of the bacteriophage study to be published (Bergan, *manuscripts in prepara-*

tion), the computing time would be 1.7×10^{22} years! Gower (27) suggested an approximation of the above method (17), but he could not overcome its principal disadvantages. Although establishing a hierarchy, divisive methods lead to less homogeneous clusters than the agglomerative procedures. Because progressively larger numbers of characters are simply excluded from comparison after each clustering cycle, divisive procedures violate Adansonian principles; still, such techniques may be useful for constructing diagnostic keys (26, 28). Of assistance in the establishment of such keys is, intuitively, also the probability of character occurrence within the cluster. Accordingly, Beers and Lockhart (5) introduced a formal statistic $P = n_+/n_c$, where n_+ is the number of positive responses among a total of n_c ONU's within the cluster. The "hypothetical median organism" (47) or the "hypothetical mean organism" (HMO) procedure (84) would also be useful tools in constructing diagnostic keys. In HMO, after completion of the clustering, the most frequent characters within clusters are listed until their number equals the mean number of positive characters in each particular phenon. Also related is consideration of the number of characters which efficiently separate single bacterial strains (30, 69). Further numerical procedures for achieving diagnostic keys are discussed extensively elsewhere (29, 59, 60).

Ward (85) described an approach, "hierarchical grouping," whereby the grouping proceeded not from highest affinities, as is customary, but with the lowest similarities. Group fusions must minimize the squared deviations about the group mean, and an "objective-function" value, $Z[i, j, k - 1]$, indicated goodness for the "inverted" fusion.

Miscellaneous Techniques

Somewhat related to clustering methods, but actually nonhierarchical, are procedures used in "operational research" (28). Here, the chief concern is the achievement of optimal multipoint set configuration, for instance, the task of finding the shortest composite distance between a set of towns. Solutions to these problems may be visualized by "minimum spanning trees" (28).

Another, widely different approach used to create correlative group structure is "factor analysis," employing vector calculus; this is mathematically somewhat more involved. Detailed presentations of the mathematical formulations for procedures related to factor analysis (32, 66, 77), "principal components analysis" (26, 37, 63; Boyce, Ph.D. thesis, Univ. of Oxford, 1963), and "discriminant analysis" (18) are outside the scope of this presentation.

Controls on the Results of Classification

Interconnected with numerical allocating procedures is the question of the goodness with which the resulting group structure reflects the original data matrix. Unfortunately, no exact procedure for such testing exists. Frequently, in taxonomy, comparisons of empirical and numerical classifications have been surprisingly similar. Results of the sorting procedures may, however, be tested by various correlation procedures. Sokal and Rohlf (80) introduced the cophenetic correlation tests whereby dendrograms were compared with each other and with the similarity matrix on which they rest. Some conclusions from such techniques will be discussed elsewhere (Bergan, *manuscripts in preparation*). Hypothesis testing is another possible approach, but the validity of *t* tests used previously (38, 83) for these types of problems is doubtful since the elements are probably not normally distributed. It is also questionable that the elements are stochastic.

The success of clustering may, furthermore, be analyzed by entropy calculations (*vide supra*). Gyllenberg (29) employed geometric theory representing heterogeneity by the radius $r = 2(\sum d^2/n_c)^{1/2}$, where d equals the distance between each ONU and the centroid within the cluster, in which n_c is the number of ONU's.

CONCLUDING REMARKS

The most recent wider scope evaluation of cluster procedures has been presented by Sneath (76). With the considerable attention presently rendered to numerical grouping procedures, a significant volume of work is annually added to the literature. Contributions to this field appear in a highly varied assortment of publications from unrelated fields of biology and technology. Unavoidably, therefore, recent but relevant developments which deserve mention may have been excluded from this survey. Nevertheless, this review should render a brief presentation of the variety of considerations involved and inherently indicate the motivation for the choice of the UWPGA and the WPGA for the numerical grouping of typing bacteriophages of *Pseudomonas aeruginosa* according to their lytic spectra (Bergan, *manuscripts in preparation*).

LITERATURE CITED

- Ball, G. H. 1965. Data analysis in the social sciences: what about the details, p. 533-559. Proc. Fall Joint Computer Conference.
- Ball, G. H., and D. I. Hall. 1965. Isodata, a novel method of data analysis and pattern classification. Stanford Research Institute, Calif.
- Baron, D. N., and P. M. Fraser. 1968. Medical applications of taxonomic methods. Brit. Med. Bull. 24:236-240.
- Beers, R. J., J. Fisher, S. Megraw, and W. R. Lockhart. 1962. A comparison of methods for computer taxonomy. J. Gen. Microbiol. 28:641-652.
- Beers, R. J., and W. R. Lockhart. 1962. Experimental methods in computer taxonomy. J. Gen. Microbiol. 28:633-640.
- Bhattacharyya, A. 1945/46. On a measure of divergence between two multinomial populations. Sankhya 7:401-406.
- Bonner, R. E. 1964. On some clustering techniques. IBM J. Res. Develop. 8:22-32.
- Brisbane, P. G., and A. D. Rovira. 1961. A comparison of methods for classifying rhizosphere bacteria. J. Gen. Microbiol. 26:379-392.
- Cain, A. J., and G. A. Harrison. 1958. An analysis of the taxonomist's judgement of affinity. Proc. Zool. Soc. (London) 131:85-98.
- Cattell, R. 1944. A note on correlation clusters and cluster search methods. Psychometrika 9:169-184.
- Cole, A. J. (ed.). 1969. Numerical taxonomy. Academic Press Inc., New York.
- Cole, L. C. 1949. The measurement of interspecific association. Ecology 30:411-424.
- Colman, G. 1968. The application of computers to the classification of streptococci. J. Gen. Microbiol. 50:149-158.
- Davis, G. H. G., and K. G. Newton. 1969. Numerical taxonomy of some named coryneform bacteria. J. Gen. Microbiol. 56:195-214.
- Dice, L. R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297-302.
- Döving, K. B. 1970. Experiments in olfaction, p. 197-225. In G. E. W. Wolstenholme and J. Knight (ed.), Ciba Symposium on Taste and Smell in Vertebrates. J. & A. Churchill. London.
- Edwards, A. W. F., and L. L. Cavalli-Sforza. 1965. A method for cluster analysis. Biometrics 21:362-375.
- Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. (London) 7:179-188.
- Flake, R. H., and B. L. Turner. 1968. Numerical classification for taxonomic problems. J. Theor. Biol. 20:260-270.
- Focht, D. D., and W. R. Lockhart. 1965. Numerical survey of some bacterial taxa. J. Bacteriol. 90:1314-1319.
- Forgey, E. W. 1964. Evaluation of several methods for detecting sample mixtures from different N-dimensional populations. American Psychological Association, Los Angeles, Calif.
- Ghent, A. W. 1963. Kendall's "tau" coefficient as an index of similarity in comparisons of plant or animal communities. Can. Entomol. 95:568-575.
- Goodall, D. W. 1964. A probabilistic similarity index. Nature (London) 203:1098.
- Goodall, D. W. 1966. A new similarity index based on probability. Biometrics 22:882-907.
- Goodall, D. W. 1966. Hypothesis testing in classification. Nature (London) 211:329-330.
- Gower, J. C. 1966. Multivariate analysis and multidimensional geometry. The Statistician 17:13-28.
- Gower, J. C. 1967. A comparison of some methods of cluster analysis. Biometrics 23:623-637.
- Gower, J. C. 1969. A survey of numerical methods useful in taxonomy. Acarologia 11:357-375.
- Gyllenberg, H. G. 1965. A model for computer identification of microorganisms. J. Gen. Microbiol. 39:401-405.
- Gyllenberg, H. G.: A general method for deriving determinative schemes for random collections of microbial isolates. Ann. Acad. Sci. Fenn. Ser. A, IV. Biologica, No. 69, 1-23.
- Hall, A. V. 1967. Methods for demonstrating resemblance in taxonomy and ecology. Nature (London) 214:830-831.
- Harmann, H. H. 1960. Modern factor analysis. The University of Chicago Press, Chicago.
- Hayhoe, F. G. J., D. Quaglino, and R. Doll. 1964. The cytology and cytochemistry of acute leukemias. Medical Research Council Special Report Series, no. 304. HMSO. London.
- Hill, L. R., M. Turri, E. Gilardi, and L. G. Silvestri. 1961.

- Quantitative methods in the systematics of *Actinomycetales*. II. G. Microbiol. 9:56-72.
35. Hill, L. R., L. G. Silvestri, P. Ihm, G. Farchi, and P. Lanciani. 1965. Automatic classification of staphylococci by principal-component analysis and a gradient method. J. Bacteriol. 89:1393-1401.
 36. Hodson, F. R., P. H. A. Sneath, and J. E. Doran. 1966. Some experiments in the numerical analysis of archaeological data. Biometrika 53:311-324.
 37. Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 26:417-441.
 38. Hubálek, Z. 1969. Numerical taxonomy of genera *Micrococcus* Cohn and *Sarcina* Goodsir. J. Gen. Microbiol. 57:349-363.
 39. Hyvärinen, L. 1962. Classification of qualitative data. B.I.T. 2:83-89.
 40. Jaccard, P. 1901. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. Bull. Soc. Vaudoise. Sci. Natur. 37:241-272.
 41. Jancey, R. C. 1966. Multidimensional group analysis. Aust. J. Bot. 14:127-130.
 42. Johnson, S. C. 1967. Hierarchical clustering schemes. Psychometrika 32:241-254.
 43. Kendrick, W. B., and J. R. Proctor. 1964. Computer taxonomy in the fungi imperfecti. Can. J. Bot. 42:65-87.
 44. Kulczynski, S. 1927. Die Pflanzenassoziationen der Pieninen. Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat. Ser. B. Suppl. II, p. 57-203.
 45. Lance, G. N., and W. T. Williams. 1966. A generalized sorting strategy for computer classifications. Nature (London) 212:218.
 46. Lessel, E. F., and J. G. Holt. 1970. Presenting and interpreting the results, p. 50-58. In W. R. Lockhart and J. Liston (ed.), Methods for numerical taxonomy. American Society for Microbiology, Bethesda, Md.
 47. Liston, J., W. Wiebe, and R. R. Colwell. 1963. Quantitative approach to the study of bacterial species. J. Bacteriol. 85:1061-1070.
 48. Lockhart, W. R. 1970. Coding the data, p. 22-33. In W. R. Lockhart and J. Liston (ed.), Methods for numerical taxonomy. American Society for Microbiology, Bethesda, Md.
 49. Lockhart, W. R., and P. A. Hartman. 1963. Formation of monothetic groups in quantitative bacterial taxonomy. J. Bacteriol. 85:68-77.
 50. MacQueen, J. B. 1966. Some methods for classification and analysis of multivariate observations. Western Management Sci. Inst. Univ. of California Working Paper no. 96.
 51. Mahalanobis, P. C., D. N. Majumdar, and C. R. Rao. 1949. Anthropometric survey of the United Provinces, 1941: a statistical study. Sankhya 9:89-324.
 52. t'Mannetje, L. 1967. A comparison of eight numerical procedures applied to the classification of some African *Trifolium* taxa based on *Rhizobium* affinities. Aust. J. Bot. 15:521-528.
 53. t'Mannetje, L. 1967. A re-examination of the taxonomy of the genus *Rhizobium* and related genera using numerical analysis. Antonie v. Leeuwenhoek J. Microbiol. Serol. 33: 477-491.
 54. Mayr, E. 1965. Numerical phenetics and taxonomic theory. Syst. Zool. 14:73-97.
 55. McQuitty, L. L. 1960. Hierarchical linkage analysis for the isolation of types. Educ. Psychol. Measurement 20:55-67.
 56. Michener, C. D., and R. R. Sokal. 1957. A quantitative approach to a problem in classification. Evolution 11:130-162.
 57. Morishima, H. and H.-I. Oka. 1960. The pattern of interspecific variation in the genus *Oryza*: its quantitative representation by statistical methods. Evolution 14:153-165.
 58. Needham, R. M. 1967. Automatic classification in linguistics. The statistician 17:45-54.
 59. Niemelä, S. I., and H. G. Gyllenberg. 1968. Application of numerical methods to the identification of micro-organisms. Folia Fac. Sci. Univ. Brunensis 7: Ser. K43, p. 279-289.
 60. Niemelä, S. I., J. W. Hopkins, and C. Quadling. 1968. Selecting an economical binary test battery for a set of microbial cultures. Can. J. Microbiol. 14:271-279.
 61. Nigel da Silva, G. A., and J. G. Holt. 1965. Numerical taxonomy of certain coryneform bacteria. J. Bacteriol. 90:921-927.
 62. Prim, R. C. 1957. Shortest connection networks and some generalizations. Bell Syst. Tech. J. 36:1389-1401.
 63. Quadling, C. 1967. Evaluation of tests and grouping of cultures by a two-stage principal component method. Can. J. Microbiol. 13:1379-1395.
 64. Quadling, C. 1970. Analyzing the data, p. 34-49. In W. R. Lockhart and J. Liston (ed.), Methods for numerical taxonomy. American Society for Microbiology, Bethesda, Md.
 65. Rogers, D. J., and T. T. Tanimoto. 1960. A computer program for classifying plants. Science 132:1115-1118.
 66. Rohlf, F. J., and R. R. Sokal. 1962. The description of taxonomic relationships by factor analysis. Syst. Zool. 11:1-16.
 67. Rohlf, F. J., and R. R. Sokal. 1965. Coefficients of correlation and distance in numerical taxonomy. Kans. Univ. Sci. Bull. 45:3-27.
 68. Rubin, J. 1967. Optimal classification into groups: an approach for solving the taxonomy problem. J. Theoret. Biol. 15:103-144.
 69. Rypka, E. W., W. E. Clapper, I. G. Bowen, and R. Babb. 1967. A model for the identification of bacteria. J. Gen. Microbiol. 46:407-424.
 70. Sebestyen, G. S. 1962. Pattern recognition by an adaptive process of sample set construction. IRE Trans. Inf. Theory IT-8:82-91.
 71. Shepard, R. N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. I. Psychometrika 27:125-140.
 72. Shepard, R. N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. II. Psychometrika 27:219-246.
 73. Sneath, P. H. A. 1957. The application of computers to taxonomy. J. Gen. Microbiol. 17:201-226.
 74. Sneath, P. H. A. 1962. The construction of taxonomic groups, p. 289-332. In G. C. Ainsworth and P. H. A. Sneath (ed.), Microbial classification. 12th Symposium of the Society for General Microbiol. Cambridge Univ. Press, Cambridge.
 75. Sneath, P. H. A. 1964. New approaches to bacterial taxonomy: use of computers. Annu. Rev. Microbiol. 18:335-346.
 76. Sneath, P. H. A. 1969. Evaluation of clustering methods, p. 257-267. In A. J. Cole (ed.), Numerical taxonomy. Academic Press Inc., New York.
 77. Sokal, R. R. 1958. Thurstone's analytical method for simple structure and a mass modification thereof. Psychometrika 23:237-257.
 78. Sokal, R. R. 1961. Distance as a measure of taxonomic similarity. Syst. Zool. 10:70-79.
 79. Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Kans. Univ. Sci. Bull. 38:1409-1438.
 80. Sokal, R. R., and F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxonomy 11:33-40.
 81. Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco.
 82. Sörensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on

- Danish commons. Kgl. Danske Vitensk. Selsk. Biol. Skr. 5(4):1-34.
83. Tsukamura, M. 1967. A statistical approach to the definition of bacterial species. Jap. J. Microbiol. 11:213-220.
84. Tsukamura, M. 1969. Numerical taxonomy of the genus *Nocardia*. J. Gen. Microbiol. 56:265-287.
85. Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Ass. 58:236-244.
86. Weber, E. 1964. Grundriss der biologischen Statistik, 5th ed. VEB. Gustav Fischer Verlag, Jena.
87. Williams, W. T., and J. M. Lambert. 1959. Multivariate methods in plant ecology. I. Association analysis in plant communities. J. Ecol. 47:83-101.
88. Williams, W. T., J. M. Lambert, and G. N. Lance. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. J. Ecol. 54:427-445.
89. Wishart, D. 1969. Mode analysis: a generalization of nearest neighbour which reduces chaining effects, p. 282-308. In A. J. Cole (ed.), Numerical taxonomy. Academic Press Inc., New York.